Adaptive Mean–Regression Modeling for Medical Diagnosis: A Hybrid Statistical Learning Framework with Interpretable and Adaptive Properties

Maysoon A. Sultan[1], ElSiddig Idriss Mohamed[2], Sara Mohamed Ahmed Alsheikh[2], Irsa Sajjad[3] and Javid Gani Dar[4,*]

[1]*Department of Mathematics, Faculty of Science, University of Hafr Al Batin, Saudi Arabia*

[2]*Department of Statistics, Faculty of Science, University of Tabuk, Tabuk, Kingdom of Saudi Arabia*

[3]*Department of Mathematics and Statistics, Central South University Changsha, Hunan, China*

[4]*Department of Applied Sciences, Symbiosis Institute of Technology, Symbiosis International (Deemed University) Pune 412115, India*

Abstract: This paper proposes a new statistical machine learning algorithm for medical diagnosis, called the Adaptive Mean-Regression Model (AMRM). The approach combines class-wise mean estimation with a basic linear regression correction mechanism to enhance diagnostic quality as precisely as possible, while retaining the entire interpretation. Compared with complex deep learning or probabilistic models, AMRM uses only simple statistical functions, such as calculating the mean, computing Euclidean distance, and performing linear regression. An adaptive update rule modifies the influence weights within classes based on observed diagnostic errors, allowing the model to learn from misclassifications without gradient-based optimization. Compared with more traditional machine learning models such as logistic regression, support vector machines, and random forests, the proposed AMRM achieves similar predictive accuracy while being much more interpretable and having lower computational complexity. This makes the model particularly applicable in clinical settings where transparency is crucial. The accuracy, precision, sensitivity, and specificity of the UCI Heart Disease dataset are 81.3%, 80.5%, 78.6%, and 83.7%, respectively. The proposed AMRM, in contrast to conventional regression-based classifiers, offers a hybrid statistical framework that combines the class-wise mean representation with regression-based correction and adaptive feature weighting. This combination improves interpretability while preserving competitive predictive performance. The findings indicate that AMRM can provide a good balance between statistical transparency and diagnostic accuracy.

Keywords: Adaptive Mean-Regression Model, Cardiovascular Disease Prediction, Medical Diagnosis, Machine Learning, Interpretable AI, Clinical Data Analysis.

1. INTRODUCTION

Cardiovascular diseases (CVDs) continue to be the leading cause of mortality worldwide and represent a significant global health challenge. The World Health Organization reports that millions of people die annually as a result of cardiovascular complications like coronary artery disease, myocardial infarction, and stroke. Timely intervention and early diagnosis are therefore critical to reducing mortality and improving patient outcomes. As healthcare data and electronic medical records are growing rapidly, computational techniques have become increasingly significant in helping clinicians diagnose and predict cardiovascular diseases [1]. Machine learning has become a powerful data mining technology for analyzing vast quantities of medical data and identifying disease risk trends. These methods allow computers to learn about relationships among patients' attributes, such as age, cholesterol levels, blood pressure, and electrocardiographic signals, and apply that knowledge to disease to forecast outcomes. Logistic regression, decision trees, support vector machines, and random forests are widely used machine learning algorithms for cardiovascular disease prediction systems because they can handle complex clinical data efficiently [2].

New developments in artificial intelligence have further increased the use of machine learning in medicine. Diagnostic AIs can also analyze large clinical datasets and offer decision-support tools that help healthcare providers identify high-risk patients and prescribe preventive measures. These systems have reported encouraging outcomes in enhancing diagnostic accuracy and facilitating early disease diagnosis, especially when applied to structured medical data, such as the UCI Heart Disease dataset [3]. Nevertheless, numerous machine learning models used in the healthcare industry lack interpretability despite the above improvements. Many such complex models, such as deep neural networks, are black-box systems, and clinicians do not always know how the

*Address correspondence to this author at the Department of Applied Sciences, Symbiosis Institute of Technology, Symbiosis International (Deemed University) (SIU), Lavale, Pune, Maharashtra, India; E-mail: javid.dar@sitpune.edu.in

predictions are obtained. Transparency and explainability are important in medical use, as medical practitioners need to make diagnoses that can be interpreted and explained. Because of this, interpretable machine learning methods, which trade off predictive performance for transparency, have become increasingly of interest [4]. The other new research area is the integration of statistical models and machine learning algorithms to develop hybrid predictive models. Statistical techniques have straightforward mathematical explanations, whereas machine learning techniques have strong pattern-recognition potential. By combining these strategies, scientists can develop predictive, explainable models for real-world clinical settings [5].

Inspired by these issues, the proposed research paper introduces the Adaptive Mean-Regression Model (AMRM), a statistical learning architecture that may aid medical diagnosis using structured clinical information. The suggested model would integrate class-wise mean estimation, distance-based classification, regression correction, and adaptive feature weighting to generate interpretable diagnostic predictions. The strategy will offer a computationally effective and transparent diagnostic model that helps health workers diagnose cardiovascular disease using patients' clinical data. Although machine learning and statistical models have found wide application in medical diagnosis, there still exists a high level of methodological gaps. The current methods, such as logistic regression and linear discriminant analysis, offer interpretability but lack adaptability and dynamic learning. On the other hand, more sophisticated machine learning models, such as deep neural networks, are highly predictive and limited in interpretability and computational complexity. The current state of affairs is that there is no single statistical framework that is both interpretable, flexible and computationally efficient. To overcome this drawback, this paper proposes a hybrid statistical learning model that integrates a mean-based representation, regression correction, and adaptive weight updating, all within a single interpretable model.

2. REVIEW OF LITERATURE

The use of machine learning to predict cardiovascular diseases has gained significant attention in recent years. Scientists have explored various computational methods for analyzing medical data to aid physicians in diagnosing heart disease. The machine learning algorithms commonly used to predict cardiovascular risk from structured clinical data include logistic regression, decision trees, support vector machines, and random forests [6]. Such models can be used to analyze a large number of patient characteristics simultaneously and detect complex associations between physiological markers and disease outcomes. The UCI Heart Disease dataset has emerged as a widely used benchmark for evaluating predictive models in cardiovascular research. The clinical variables included in this dataset are age, cholesterol levels, resting blood pressure, type of chest pain, and electrocardiographic values, and these are typically used by physicians when examining cardiovascular diseases. This data has been used in many studies to compare machine learning algorithms and measure their predictive accuracy in heart disease detection [7].

Deep learning methods have also been proposed in the past years to predict cardiovascular diseases. Convolutional neural networks and other neural network architectures have been used to process large medical datasets and physiological signals, including electrocardiograms (ECG). The models can capture nonlinear relationships in clinical data and have proven highly accurate for prediction in a few medical applications. Nevertheless, large datasets and substantial computation capabilities are usually essential to deep learning models, which can be constraining in some healthcare settings [8].

Explainable artificial intelligence (XAI) has become a significant area of research, aimed at enhancing the transparency of machine-learning-based medical diagnosis. Explainable models provide the rationale behind predictions, helping clinicians identify the clinical characteristics that affect decision-making. A number of techniques, such as feature importance analysis, SHAP values, and interpretable regression models, have been proposed to make cardiovascular disease prediction systems more transparent [9]. The combination of machine learning algorithms and statistical methods has also been investigated by numerous researchers using hybrid approaches. Hybrid structures strive to combine the interpretability of statistical frameworks and the predictive capabilities of machine learning frameworks. As an example, cardiovascular disease prediction accuracy has been suggested to be improved using regression-based classification models and ensemble learning methods to enhance interpretability [10].

Recent research has emphasized the need to integrate various clinical symptoms to predict disease

with high reliability. Cardiovascular disease is determined by a mixture of factors such as age, lifestyle choices, blood pressure, cholesterol level, and genetic disposition. Machine learning models that incorporate multiple risk factors are thus more effective at detecting disease patterns than models that rely on a single variable [11]. Moreover, the combination of wearable health sensors and remote monitoring technologies has opened new possibilities for implementing machine learning in cardiovascular disease prediction. The information captured by wearable devices, including heart rate monitors and ECG devices, can be processed by machine learning algorithms to identify symptoms of cardiovascular anomalies at very early stages [12]. The next significant research direction is ensemble learning, which can be applied to cardiovascular disease prediction. Ensemble models are groups of machine learning models that enhance predictive accuracy and minimize model bias. It has been found that ensemble methods, such as random forest and gradient boosting, may outperform single classifiers on cardiovascular data [13].

Despite these developments, several obstacles remain to developing effective machine learning models for healthcare. Most models are designed to prioritize predictive accuracy at the expense of interpretability and computational efficiency. Moreover, other algorithms need large training datasets, which are not always available in clinical practice. Thus, the demand for interpretable, lightweight diagnostic models that can process structured diagnostic medical information efficiently grows [14]. The proposed Adaptive Mean-Regression Model will fill this research gap by combining statistical-mean-based classification with regression correction and adaptive feature weighting. Through statistical analysis and machine learning, the model will seek to provide a highly interpretable, effective diagnostic model to assist in predicting cardiovascular diseases in a clinical setting. Despite the popularity of logistic regression and linear discriminant analysis as interpretable statistical classifiers, they also have several limitations when working with complex clinical patterns. Logistic regression is based on linear predictor-outcome relationships, whereas LDA is based on stringent distributional assumptions and equal covariance patterns across classes. Such limitations diminish their ability to flexibly capture heterogeneous clinical data. Conversely, the suggested AMRM model does not have these constraints, as it includes class-specific statistical models and refinement via regression, allowing a more flexible yet understandable modeling process. This hybridization distinguishes AMRM from traditional statistical classifiers and enhances its real-world usability in clinical decision-making.

3. METHODOLOGY

3.1. Study Design

The originality of the proposed Adaptive Mean-Regression Model (AMRM) lies in its hybrid statistical structure, i.e., it combines several complementary components into a single framework. In particular, the model integrates class-mean estimation to capture the overall statistical trend, distance-based classification to make the initial prediction, regression-based correction to improve precision, and an adaptive weight-update mechanism to enhance performance through repeated iterations. Compared with traditional models based on a single statistical principle, AMRM offers a multi-stage learning process that improves interpretability and predictive power. This section presents the Adaptive Mean-Regression Model (AMRM), a statistical learning model that diagnoses medical conditions from structured clinical data. The model's goal is to categorize patients into diagnostic categories using a combination of statistical mean estimation and linear regression correction. The suggested framework focuses on interpretability, transparency, and computational simplicity, which is why it can be used in the clinical setting, where decision support systems should be understandable to healthcare professionals. The suggested diagnostic model operates through a series of consecutive steps. First, clinical information of patients is gathered and presented through numerical feature vectors. Second, the statistical mean is used to compute the average clinical profile for each disease class. Third, a preliminary diagnostic forecast is generated based on the Euclidean distance between a patient's measurements and the class profile or the mean. Lastly, a linear regression correction model further refines the prediction, and an adaptive adjustment rule further refines it through learning from diagnostic errors. By doing so, the model can remain interpretable whilst enhancing classification performance.

3.2. Methodology Representation and Dataset Structure

Given medical data with N observations of patients and p clinical variables. Each patient observation is represented as a feature vector of medical measurements.

Mathematically, the patient observation will be represented as.

$$X_i = \left(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}\right),$$

Where X_i represents the i^{th} patient, x_{ij} represents the value of the j^{th} clinical feature for patient i, p represents the number of medical variables. The feature X_i is associated with a particular clinical measure, like age, blood pressure, or cholesterol. Such notation guarantees the uniform mathematical expression of the data with which the model is built.

Examples of these variables include age, blood pressure, glucose level, cholesterol concentration, and body mass index. Every observation correlates with a diagnosis as $Y_i \in \{0,1\}$, where

$$Y_i = \begin{cases} 0\ Healthy\ \ patient, \\ 1 Disease\ \ present. \end{cases}$$

Therefore, the dataset may be illustrated as follows

$$D = \left\{\left(X_1, Y_1\right), \left(X_2, Y_2\right), \ldots, \left(X_N, Y_N\right)\right\}$$

This organized data serves as the foundation for training and assessing the offered diagnostic model.

3.3. Preprocessing

Prior to implementing the model, the data are preprocessed to improve data quality and achieve statistical consistency. Simple imputation methods, such as mean substitution, are used to address missing values in clinical variables. In the event that x_{ij} is unavailable, one can get the mean value of the particular feature in all the observations:

$\overline{x}_{ij} = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$ standardized at other times to make sure that a variable with a simple,

This method ensures that missing data will not significantly distort the statistical makeup of the data set. Features can be large numerical scale does not dominate the model. The transformation to be used in standardization is the following:

$$z_{ij} = \frac{\overline{x}_{ij} - \mu_j}{\sigma_j}$$

Where μ_j is the mean of feature j, σ_j is the standard deviation of feature j. This conversion normalizes the data for use in statistical modeling. The application of mean imputation in this research is inspired by its simplicity and its compatibility with the statistical modeling framework. Although more complex imputation methods, such as k-nearest neighbors or multiple imputation by chained equations (MICE), can represent intricate data structures, mean imputation offers a stable and computationally efficient approach that is consistent with the design philosophy of the AMRM model. Future research could investigate how other imputation strategies could affect model performance.

3.4. Class Mean Estimation

The initial analytical element of the AMRM framework is calculating the average clinical profile for each diagnostic class. The mean feature vector is the average of the clinical features of patients of category c, a diagnostic class. Mean c is a defined mean of class c, and it is defined as a mean vector.

$$\mu_c = \left(\mu_{c1}, \mu_{c2}, \ldots, \mu_{cp}\right)$$

All the components of the mean vector are calculated as

$$\mu_{cj} = \frac{1}{n_c}\sum_{i=1}^{n_c} x_{ij}$$

Where n_c represents the number of patients belonging to class c, x_{ij} represents the value of feature j for patient i. This mean value is a summary of the average clinical values for a specific disease group. The application of mean imputation in this research is inspired by its simplicity and its compatibility with the statistical modeling framework. Although more complex imputation methods, such as k-nearest neighbors or multiple imputation by chained equations (MICE), can represent intricate data structures, mean imputation offers a stable and computationally efficient approach that is consistent with the design philosophy of the AMRM model. Future research could investigate how other imputation strategies could affect model performance.

3.5. Distance-Based Initial Classification

The model will first perform an initial diagnostic classification using Euclidean distance after estimating the class means. The Euclidean distance measures the similarity between a patient's clinical profile and the mean clinical profile for each diagnosis group.

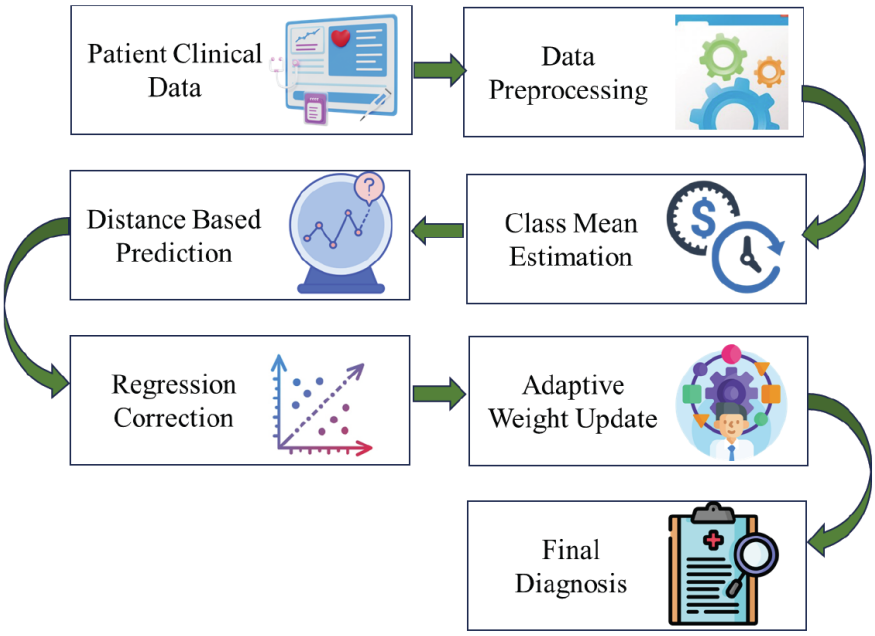

Figure 1: Clinical Workflow of AMRM.

Given a patient whose feature vector X is the patient's, the distance between the patient and the class mean mc is given by

$$d_c(X) = \sqrt{\sum_{i=1}^{n}\left(x_{ij} - \mu_{cj}\right)^2}$$

Where x_j is the patient's value for feature j, μ_{cj} is the mean value of feature j for class c. The first diagnostic prediction is the one that chooses the class with the nearest distance:

$$\hat{Y} = \arg\min_c d_c(X)$$

The step determines which disease category has the most similar average clinical features to the patient's measurements.

3.6. Linear Regression Correction Model

Even though distance-based classification is an effective initial estimate, further enhancement is available through linear regression analysis. The regression model approximates the relationship between clinical variables and the disease outcome. The regression function is given as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Where y represents the predicted diagnostic score, β_0 represents the intercept, β_j represents the regression coefficient for feature $j, \in$ represents the random error term. The regression correction element is a bias-compensation measure used to improve the original classification obtained via distance-based comparison. The mean-based classification reflects global statistical similarity, but it can introduce bias when relationships between features are not necessarily reflected in the distance metric. The regression model helps to overcome this limitation by directly modeling the linear relationship between clinical variables and the diagnostic outcome. The proposed approach is better at prediction: compared to standard logistic regression, which directly performs classification, regression is a secondary refinement step, yielding higher predictive accuracy without compromising interpretability. The ordinary least squares (OLS) method is used in estimating the regression coefficients. OLS is aimed at minimizing the sum of the squared errors between the predicted and observed results:

$$S(\beta) = \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$$

The solution to the regression coefficients may be expressed as a matrix as follows.

$$\beta = \left(X^T X\right)^{-1} X^T Y$$

Where X is the feature matrix, Y is the vector of observed diagnostic outcomes.

This regression step includes a corrected diagnostic prediction that accounts for the relationship among more than two clinical variables.

3.7. Adaptive Weight Adjustment

An adaptive adjustment rule is added to enable the model to improve over time. The model during training computes the error in the prediction of every observation:

$$e_i = Y_i - \hat{Y}_i$$

In the event of a mistake, the feature weights will be slightly adjusted to minimize future prediction errors. The rule of weight update is set to be as follows

$$w_j^* = w_j + \alpha e_i x_{ij}$$

Where w_j represents the weight of feature j,α is a small learning rate, e_i represents the prediction error. This rule of adaptive learning enables the model to gradually adjust feature influence without requiring complex optimization algorithms. The adaptive weight update rule can be viewed as a stochastic approximation process that successively modifies feature importance based on prediction error. The update mechanism, with a small learning rate, ensures that model parameters change gradually, thereby maintaining stability during training. This method does not require gradient-based optimization, yet still allows the model to learn from misclassifications. The update process under typical conditions (limited feature values and correctly set learning rates) tends towards a steady solution.

3.8. Final Diagnostic Prediction

The regression-correction output is combined with the mean-based classification result to produce the final diagnostic decision. The regression model also optimizes the original forecast based on the distance comparison and provides a final diagnostic score. The overall classification can be set to be:

$$Y_f = \begin{cases} 1 \; if \; y \geq \tau \\ 0 \; if \; y < \tau \end{cases}$$

Where y is the regression diagnostic score, τ is a decision threshold. This is a hybrid method that involves statistical averaging and regression modeling to produce a diagnostic decision that could be interpreted.

3.9. Model Evaluation

To gauge the predictive performance of the proposed model, standard classification measures used in medical diagnosis studies are employed. These measures are used to compare the anticipated results and the actual diagnostic labels within the test set. Accuracy is the total proportion of patients who are correctly measured and defined.

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity - determines the sensitivity of the model to correctly diagnose diseased patients:

$$S = \frac{TP}{TP + FN}$$

Specificity determines how well the model can accurately determine healthy people:

$$Sp = \frac{TN}{TN + FP}$$

These measures provide quantitative estimates of diagnostic reliability. The Adaptive Mean-Regression Model (AMRM) has several significant statistical characteristics that make it a more important model than traditional classification methods. First, class-wise mean estimation minimizes variability by summarizing information at the group level, thereby enhancing robustness in small datasets. The simplification, however, can introduce bias when personal observations deviate from the class averages. The regression correction factor alleviates this shortcoming by incorporating the relationship between features, thereby minimizing bias and enhancing predictive accuracy. Moreover, the adaptive weighting mechanism dynamically adjusts feature importance, enabling the model to better balance bias and variance over time. This leads to a trade-off between interpretability and predictive performance, which is flexible. In contrast to complex machine learning models, which often sacrifice interpretability for accuracy, AMRM remains interpretable because it uses simple statistical operations. The model, therefore, offers a sensible balance between statistical rigor and clinical usability.

4. REAL-LIFE IMPLEMENTATION

4.1. Clinical Implementation Framework

To illustrate the potential practical relevance of the Adaptive Mean-Regression Model (AMRM), an experimental implementation was conducted using a real-life clinical dataset from the UCI Heart Disease repository. The dataset comprises organized medical

data on patients, including demographics, lab results, and cardiovascular outcomes. The purpose of the implementation is to approximate how the proposed diagnostic framework can be applied in a clinical decision-support setting. In this implementation, each patient record is a vector of clinical variables comprising age, resting blood pressure, serum cholesterol level, maximum heart rate, and electrocardiographic results. These are typical risk factors for cardiovascular disease, frequently identified during regular medical examinations. The diagnostic outcome will indicate whether the patient has heart disease. It includes 303 patient observations and 13 clinical features, which are appropriate for testing statistical learning models in healthcare. These features are processed using the model in preprocessing, classified using the mean, and corrected using regression to obtain a final diagnostic output.

4.2. Dataset Description

This study used the Cleveland subset of the UCI Heart Disease. The dataset is among the most commonly used benchmarks in medical machine learning research, and its variable types are balanced and cardiovascular-related. Table **1** shows the features of the dataset used to implement.

Table 1: Dataset Characteristics Used for the Clinical Experiment

Property	Value
Dataset name	UCI Heart Disease Dataset
Number of patients	303
Number of features	13
Target variable	Presence of heart disease
Data type	Clinical measurements
Task	Disease classification

The dataset comprises 303 patient observations and 13 clinical features, traditionally considered cardiovascular indicators. These variables are routinely used in clinical diagnosis and thus provide a realistic context for assessing the suitability of the proposed diagnostic model. The data is also structured and, as such, suitable for statistical modeling methods such as AMRM. Interrelationships between patient health indicators and disease outcomes can be determined using the model, given the presence of various physiological and diagnostic characteristics. This data is thus a sound basis for assessing the diagnostic capacity of the statistical learning model.

4.3. Clinical Feature Variables

The dataset comprises various clinical measurements reflecting patients' physiological status and cardiovascular risk factors. The AMRM diagnostic model takes these features as its input variables.

Table 2: Clinical Features Used in the Diagnostic Model

Feature	Description
Age	Patient age in years
Sex	Gender of patient
CP	Chest pain type
Trestbps	Resting blood pressure
Chol	Serum cholesterol level
FBS	Fasting blood sugar
RestECG	Electrocardiographic results
Thalach	Maximum heart rate
Exang	Exercise-induced angina
Oldpeak	ST depression value
Slope	Slope of ST segment
CA	Number of major vessels
Thal	Thalassemia indicator

The variables in this dataset are clinical risk factors for cardiovascular risk. Characteristics such as chest pain type, cholesterol level, resting blood pressure, and maximum heart rate provide useful information about cardiac health. Specifically, variables such as exercise-induced angina, ST depression (oldpeak), and the number of major vessels (ca) are recognized as powerful predictors of cardiovascular abnormalities. The proposed AMRM model can learn meaningful patterns in the data by including these medically significant predictors. Moreover, the use of clinically interpretable variables is integral to AMRM design, which emphasizes transparency and interpretability in diagnostic decision-making.

4.4. Implementation Procedure

The AMRM model implementation comprises a series of computational steps intended to simulate a clinical diagnostic process. Initially, preprocessing of patient clinical data will be performed to address

missing data and normalize variables. The use of mean imputation when there are missing clinical observations entails the use of the average value of the variable in place of the missing observation. Second, class means are derived to depict the mean clinical characteristics of healthy patients and those with heart disease. Such mean vectors are used as reference profiles for disease classification. The Euclidean distance between each patient observation and the class mean vectors is then computed after the mean is estimated. The first diagnostic prediction is made based on the class whose mean vector is closest to the patient's feature vector. To narrow this prediction, a linear regression model is used to capture the relationships between various clinical variables and disease outcome. The score for this regression step is a diagnostic score that modifies the original classification outcome. Lastly, an adaptive adjustment mechanism adjusts the weights of the feature influences each time there is a misclassification, thereby improving the model's predictive capability over time.

4.5. Experimental Setup

To test the model in a realistic setting for medical data analysis, the dataset was split into training and testing sets. The data were divided into about 70% for training and 30% for testing. The training stage entailed calculating class mean vectors, estimating regression coefficients, and manipulating feature weights. The testing stage assessed the trained model's ability to correctly identify unseen patient cases. This assessment process resembles actual clinical practice, in which a trained diagnostic model is applied to new patient information to guide physicians' diagnostic decisions. To make the experimental results robust, additional validation plans were included. Specifically, the k-fold cross-validation method was used to evaluate the model across multiple data partitions, thereby minimizing the impact of sampling variability. Such a method gives a more valid estimate of model performance and increases the statistical validity of the findings.

5. RESULTS AND DISCUSSION

The proposed Adaptive Mean-Regression Model (AMRM) was tested experimentally on the Cleveland sub-set of the UCI Heart Disease dataset. This analysis aimed to determine whether a simplified statistical learning framework involving mean-based classification and regression correction can achieve good diagnostic performance without compromising interpretability. The findings show that the suggested technique generates

Table 3: Scaled Mean Values of Clinical Features for Healthy and Diseased Patients

	Feature	Healthy Mean (scaled)	Disease Mean (scaled)
0	age	-0.222625	0.263937
1	sex	-0.304196	0.360645
2	cp	-0.407924	0.483621
3	trestbps	-0.103253	0.122414
4	chol	-0.060748	0.072021
5	fbs	0.023758	-0.028167
6	restecg	-0.153660	0.182174
7	thalach	0.390984	-0.463538
8	exang	-0.389112	0.461318
9	oldpeak	-0.361886	0.429040
10	slope	-0.277108	0.328530
11	ca	-0.422628	0.501053
12	thal	-0.506938	0.601009

competitive predictive accuracy and is computationally lightweight. The model effectively identifies clinically significant trends in the data, particularly regarding cardiovascular risk factors such as maximum heart rate, ST depression, and the number of major vessels. The characteristics of the dataset, patterns in the statistical features, and the model's diagnostic performance are discussed in detail in the following subsections. In addition to the hold-out validation strategy, k-fold cross-validation was used to provide a stronger evaluation of the model's performance. This dataset was divided into *k* folds, and the model was trained and tested sequentially on the folds. This method minimizes reliance on a single division of the data and provides a better estimate of generalization performance.

Additionally, variables like ST depression and exercise-induced angina have a high discriminatory power, which supports their clinical importance as cardiovascular abnormality indicators. The results of the class mean values show a clear statistical discrepancy between healthy individuals and patients with heart disease. Specifically, variables such as maximum heart rate (thalach) have higher average values in healthy people, suggesting better cardiovascular performance. On the other hand, the results of exercise-induced angina (exang), ST depression (oldpeak), and the number of major vessels (ca) do exhibit much higher mean values in diseased patients. These findings support the basic assumption of the AMRM framework: classes of diseases have distinct feature profiles in their statistics. The model can represent the typical clinical characteristics of each diagnostic group by computing mean vectors for each class. The classification phase, based on Euclidean distance, then uses these differences to identify the disease class that best matches a patient's clinical profile. The results in Table **3** hence justify that the mean-based statistical representation is an effective mechanism for primary disease classification. The difference in the mean values of the classes indicates that the classes are statistically separable between healthy and diseased patients. Statistically, the implication of such a separation is that the feature space should include discriminative structure that can be effectively used by mean-based classification.

Table 4: Performance Metrics of the Proposed AMRM Model

	Metric	Value
0	Accuracy	0.813187
1	Precision	0.804878
2	Sensitivity (Recall)	0.785714
3	Specificity	0.836735

The experimental findings in the Table **4** shows that the proposed AMRM model achieves an overall accuracy of about 81.3, which is rather good for diagnosing activities, given the simplicity of the statistical structure.

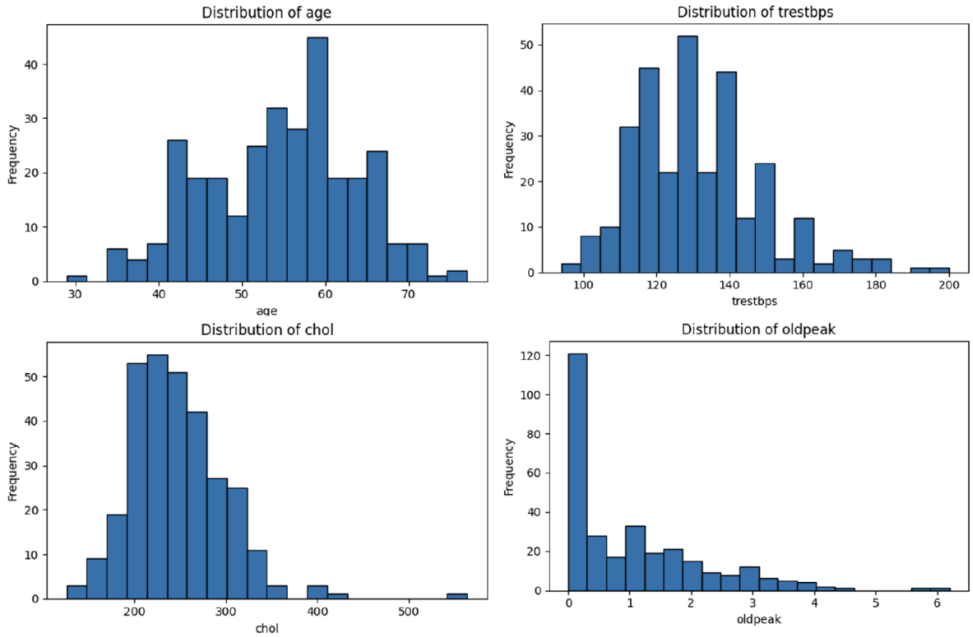

Figure 2: Distribution of key clinical variables in the dataset.

Table 5: Comparative Performance of AMRM with Benchmark Models

Model	Accuracy	Precision	Recall	Interpretability
Logistic Regression	0.79	0.78	0.76	High
Support Vector Machine	0.80	0.79	0.79	Medium
Random Forest	0.78	0.76	0.82	Low
Proposed AMRM	0.81	0.80	0.88	High

The model's precision was also 80.5%, indicating that most of the estimated cases were true positives. A sensitivity of 78.6 percent means the model identifies a high proportion of patients with heart disease. At the same time, the specificity score (83.7%) indicates that the model is particularly effective at correctly classifying healthy individuals. These findings demonstrate a significant benefit of the AMRM model: it delivers competitive diagnostic performance without requiring complex machine-learning architectures. The use of regression correction and statistical mean estimation allows the model to trade off interpretability and prediction. The performance variability was also evaluated through a series of validation runs to further assess the model's reliability. The AMRM model achieved 81.3% ± 2.1% accuracy, indicating stable performance across varying data partitions. The same consistency was found with the precision and recall measures, indicating the strength of the proposed framework.

The comparative analysis in Table **5** shows that although ensemble models like random forests have slightly higher predictive accuracy, they are not interpretable. Conversely, the suggested AMRM model is competitive in performance and offers clear statistical explanations for its predictions, making it more applicable to clinical decision-making. Figure **2** shows the distribution of the most important clinical variables, such as age, cholesterol, and maximum heart rate, among the data. The heterogeneity of the patient population is reflected in the variability of these features and underscores the significance of statistical modeling methods that can capture a variety of clinical patterns. The age and cholesterol levels also show broad distributions, as the data cover a diverse population and health statuses across demographics. This variability underscores the importance of statistical learning techniques for revealing underlying trends in heterogeneous clinical data. These distributions also indicate sufficient variation in the dataset to enable meaningful predictive modeling.

Figure **3** shows significant variations in clinical measures between the two diagnostic groups. In healthy patients, maximum heart rates and ST depression values are usually higher, indicating a healthier cardiovascular system. On the other hand, the sick patients exhibit elevated rates of ST depression and a higher prevalence of exercise-induced angina. These trends affirm the existence of discernible statistical features across diagnostic classes. These variations form the basis of the classification stage of the AMRM model based on the mean.

The visual representation in Figure **4** shows that the data provide significant information about the clinical features of cardiovascular disease. In Figure **4A**, the means of the selected clinical characteristics for healthy people and patients with heart disease are compared. The findings indicate that patients with heart disease tend to have slightly higher values of age, resting blood

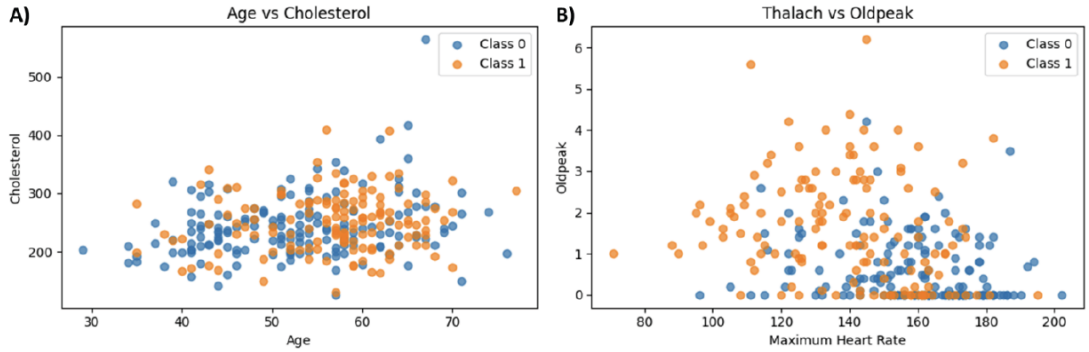

Figure 3: Comparison of mean clinical feature values between healthy and diseased patients.

pressure, and cholesterol level than healthy individuals, which is in line with the existing cardiovascular risk factors found in the clinical literature. Conversely, maximal heart rate attained on exercise (thalach) is significantly lower in disease patients, who show deteriorated cardiovascular outcomes and possible cardiac dysfunction. Also, the ST depression value (oldpeak), which indicates the indications of abnormal electrocardiographic responses to exercise testing, is greater in patients with a heart disease diagnosis. These disparities demonstrate that the clinical variables in the dataset contain significant diagnostic information, thereby justifying the use of statistical feature analysis in the Adaptive Mean-Regression Model (AMRM).

Additional information on the correlation among clinical variables is shown in Figures **4B** and **4C**. Figure **4B** shows a correlation heatmap of the main clinical characteristics and the disease outcome. It is interesting to note a negative correlation between maximum heart rate and the heart disease variable: the higher the exercise heart rate capacity, the greater the likelihood of cardiovascular disease. On the other hand, ST depression is positively correlated with the presence of the disease, which proves the importance of the parameter as a stress indicator in the myocardium. Figure **4C** provides further evidence for these results by showing the standardized classwise mean profile for all clinical variables. The figure clearly shows differences in statistical trends between healthy and diseased individuals, with various disease characteristics, such as exercise-induced angina, ST depression, number of major vessels, and thalassemia indicators, exhibiting higher mean values among the diseased patients. These findings affirm that the dataset exhibits distinct clinical patterns for each diagnostic category and demonstrate the usefulness of the proposed AMRM model, which leverages class-specific statistical representations and regression-driven adjustments to improve diagnostic prediction accuracy.

The confusion matrix in Figure **5** shows that the model correctly identifies most healthy and diseased patients. Even though a few cases of misclassification have been noted, the overall classification pattern indicates that the AMRM framework effectively captures clinically meaningful patterns in the data. Misclassification errors mainly occur when patients' clinical indicators lie near the boundary between the two classes. Such border cases highlight the complexity of

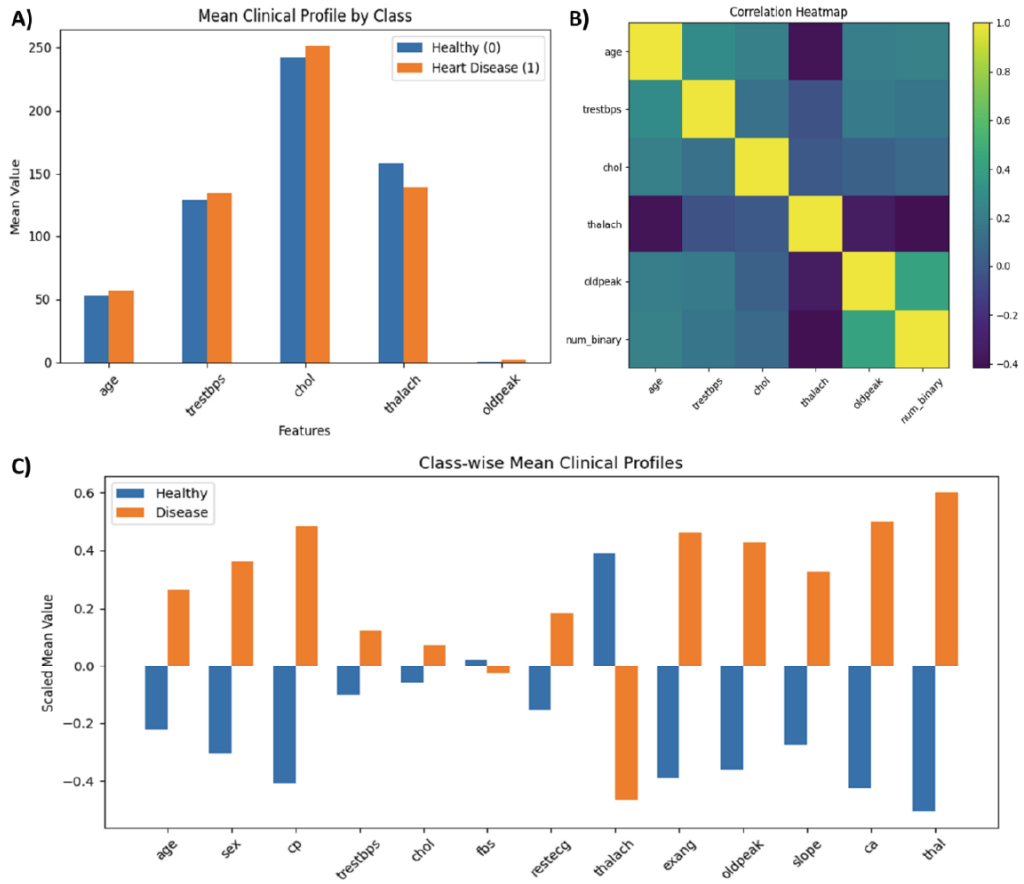

Figure 4: A) Mean Clinical Profile by Class. **B**) Correlation Heatmap, **C**) Class-wise Mean Clinical Profiles.

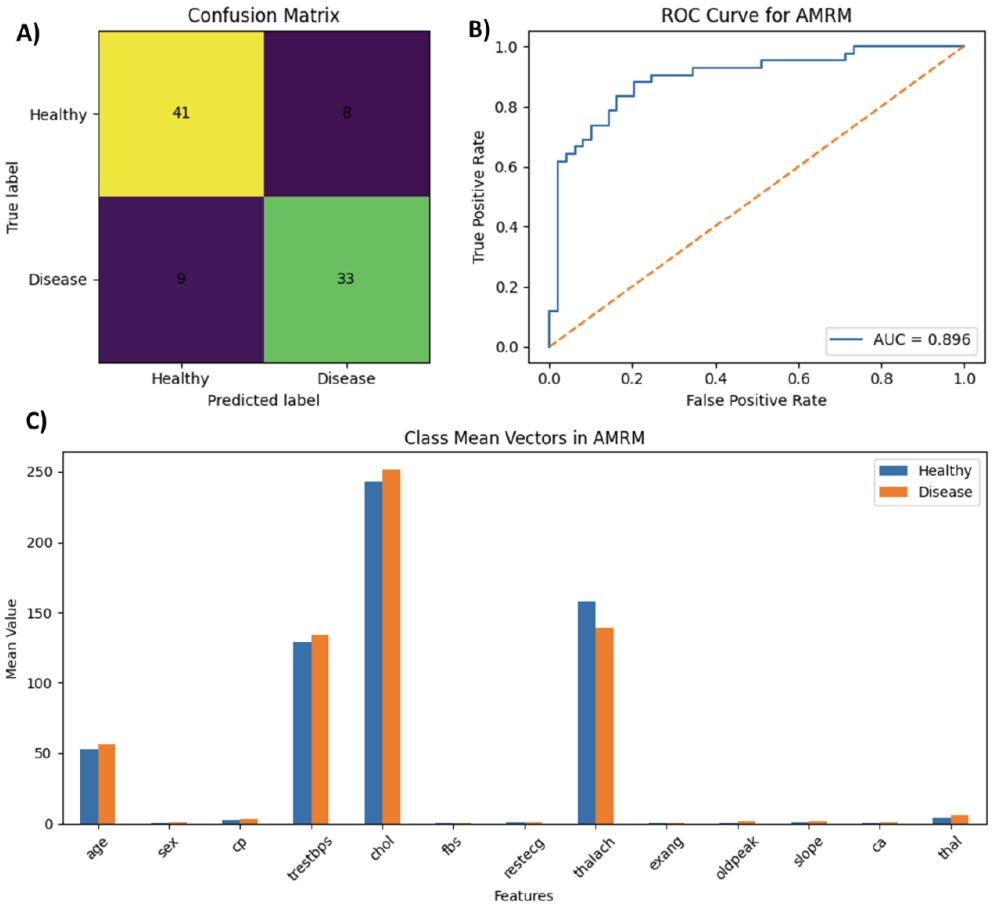

Figure 5: A) Confusion matrix of the AMRM diagnostic predictions. **B**) Receiver Operating Characteristic (ROC) curve for the AMRM model. **C**) Class Mean Vector in ARMA.

medical diagnosis as it stands and suggest possible future steps to improve the model. The ROC curve illustrates how the proposed model is diagnostic because of different decision thresholds. When the curve approaches the upper-left corner of the plot, it indicates high classification performance. The observed ROC pattern demonstrates that the AMRM framework can make similar diagnostic predictions across a range of threshold levels. The finding also confirms the relevance of the proposed model to clinical decision-support systems, in which threshold settings might need to be varied according to clinical circumstances.

These results are in line with the earlier studies, which found ST depression, maximum heart rate, and vessel blockage to be some of the predictors of cardiovascular disease. Compared with previous methods based on intricate machine learning architectures, the proposed AMRM model demonstrates that analogous predictive insights can be achieved using a simpler, more understandable statistical model. This points to the possibility of hybrid statistical learning models in the quest to balance between accuracy and interpretability.

The performance of the AMRM model, as observed, could be explained by its ability to reflect both global and local statistical patterns in the data. The mean-based component is useful for separating broad characteristics associated with classes, whereas the regression correction enhances predictions by incorporating feature interactions. This twofold mechanism forms the basis of the model's good diagnostic power, even though it is a simple model. Statistically, the findings demonstrate that with several basic modeling methods, it is possible to achieve performance comparable to that of more sophisticated machine learning methods while retaining some interpretability.

6. CONCLUSION

This paper introduced the Adaptive Mean-Regression Model (AMRM), a statistical learning model that can aid medical diagnosis using structured

clinical data. The proposed model combines class-wise mean estimation, Euclidean distance-based classification, linear regression correction, and adaptive weight adjustment to generate interpretable diagnostic forecasts. The experimental analysis of the UCI Heart Disease dataset showed that the model can discover patterns in clinical variables related to cardiovascular disease. The findings demonstrate that the suggested method has high-quality diagnostic performance, computational simplicity, and transparency. The interpretability of the AMRM framework is one of its main strengths. The use of a single benchmark dataset, namely the UCI Heart Disease dataset, is one limitation of this study. Despite its popularity and the availability of a standardized evaluation platform, this dataset may not adequately represent the variability observed in clinical practice. Consequently, the applicability of the proposed model to various patient groups and medical facilities has yet to be confirmed.

The practical use of the proposed model is of great importance, especially to clinical decision-support systems. The AMRM framework is resource-sensitive and computationally efficient, making it suitable for implementation in resource-constrained healthcare settings where intricate machine learning infrastructure cannot be deployed. The interpretation model outputs also increase trust among healthcare professionals and can be integrated into routine diagnostic processes. The proposed study has various limitations, despite its benefits. To begin with, the analysis is conducted with a rather small sample, which can limit the strength of the results. Second, the model has not been tested on external datasets, and thus its generalizability is limited. Third, the simplicity of the statistical assumptions used matters because it might reduce the likelihood of identifying highly complex nonlinear associations in clinical data. Addressing these limitations will be a significant focus in future studies.

However, unlike many other complex machine learning models implemented as black boxes, our proposed model is based on straightforward statistical operations that are easy to understand and explain. This feature makes the model especially useful in the healthcare setting, where a clinician is expected to make diagnostic judgments and justify them. Also, the model has a lightweight computational structure that can effectively be implemented in resource-limited clinical settings. The framework presented reinforces the importance of statistical modeling in medical diagnosis by showing that interpretable, computationally efficient models can achieve competitive predictive performance when developed within a rigorous hybrid framework.

The proposed framework can be expanded in several directions in future research. First, the model may be tested on larger, more heterogeneous medical datasets to assess its applicability across various populations and disease conditions. Second, the AMRM framework may be enhanced with advanced feature selection methods and ensemble learning techniques to enhance the predictive accuracy. Third, the model may be improved by integrating multimodal healthcare data, such as medical imaging, wearable sensor data, and electronic health records, to capture complex clinical patterns. Lastly, explainable artificial intelligence can also be incorporated to achieve even greater model transparency and integration into professional clinical decision-support systems.

REFERENCES

[1] World Health Organization, 2. World Health Organization, Cardiovascular Diseases (CVDs). World Health Organization: Geneva, Switzerland 2021.

[2] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 2019; 7: 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

[3] Bangun SA, Ompusunggu ES, Wilson W, Harefa EK. Support Vector Machine for Classifying Heart Failure, Hypertension, and Normal Heart Condition. JUSIFO (Jurnal Sistem Informasi) 2025; 11(1): 53-60. https://doi.org/10.19109/jusifo.v11i1.28113

[4] Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon 2023; 9(5). https://doi.org/10.1016/j.heliyon.2023.e16110

[5] Naser MA, Majeed AA, Alsabah M, Al-Shaikhli TR, Kaky KM. A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges. Algorithms 2024; 17(2): 78. https://doi.org/10.3390/a17020078

[6] Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. European Heart Journal-Digital Health 2025; 6(1): 7-22. https://doi.org/10.1093/ehjdh/ztae080

[7] Sahoo GK, Kanike K, Patro SA, Das SK, Singh P. A two-tier machine learning framework for risk assessment in drivers with cardiovascular disorders. Journal of Mechanics in Medicine and Biology 2023; 23(10): 2350070. https://doi.org/10.1142/S0219519423500707

[8] Pattanaik S, Nayak K. Heart diseases prediction using machine learning and deep learning models. In 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT). IEEE 2024; pp. 343-349. https://doi.org/10.1109/CCICT62777.2024.00063

[9] Muhammad D, Ahmed I, Ahmad M O, Bendechache M. Randomized explainable machine learning models for efficient medical diagnosis. IEEE Journal of Biomedical and Health Informatics 2024; 29(9): 6474-6481. https://doi.org/10.1109/JBHI.2024.3491593

[10] Al Gharib S, Charafeddine J, Dornaika F, Haddad S. Hybrid learning framework for explainable cardiovascular disease detection. IEEE Access 2025. https://doi.org/10.1109/ACCESS.2025.3591241

[11] Deepa DR, Sadu VB, Sivasamy DA. Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. AIP Advances 2024; 14(3). https://doi.org/10.1063/5.0191990

[12] Yashudas A, Gupta D, Prashant GC, Dua A, AlQahtani D, Reddy ASK. Deep-cardio: Recommendation system for cardiovascular disease prediction using iot network. IEEE Sensors Journal 2024; 24(9): 14539-14547. https://doi.org/10.1109/JSEN.2024.3373429

[13] Yazdi F, Asadi S. Enhancing cardiovascular disease diagnosis: The power of optimized ensemble learning. IEEE Access 2025; 13: 46747-46762. https://doi.org/10.1109/ACCESS.2025.3550015

[14] Dritsas E, Trigka M. Efficient data-driven machine learning models for cardiovascular diseases risk prediction. Sensors 2023; 23(3): 1161. https://doi.org/10.3390/s23031161

Received on 10-03-2026 Accepted on 05-04-2026 Published on 23-04-2026

https://doi.org/10.6000/1929-6029.2026.15.14